Dense Multiperson Tracking with Robust Hierarchical Linear Assignment

McLaughlin, N., Martinez-del-Rincon, J., & Miller, P. (2015). Dense Multiperson Tracking with Robust Hierarchical Linear Assignment. *IEEE Transactions on Cybernetics*, *45*(7), 1276-1288. https://doi.org/10.1109/TCYB.2014.2348314

Published in:
IEEE Transactions on Cybernetics

Document Version:
Peer reviewed version

Dense Multiperson Tracking with Robust Hierarchical Linear Assignment

Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller

***Abstract*—We introduce a novel dual-stage algorithm for online multi-target tracking in realistic conditions. In the first stage, the problem of data association between tracklets and detections, given partial occlusion, is addressed using a novel occlusion robust appearance similarity method. This is used to robustly link tracklets with detections without requiring explicit knowledge of the occluded regions. In the second stage, tracklets are linked using a novel method of constraining the linking process that removes the need for ad-hoc tracklet linking rules. In this method, links between tracklets are permitted based on their agreement with optical flow evidence. Tests of this new tracking system have been performed using several public datasets.**

***Index Terms*—Multi-Person tracking, Tracking-by-Detection, Occlusion Modelling, Surveillance, Linear Assignment**

I. Introduction

Multi-target tracking is an important component of various applications in computer vision such as visual surveillance and sports analysis. However, despite its crucial role, consistently and accurately tracking multiple interacting people over time remains a challenge. This is due to the many sources of uncertainty, e.g., measurement noise, background clutter, changing background and illumination conditions, significant occlusions and distractors. The tracking problem is made more difficult by the need to track several interacting targets with similar appearance, in overcrowded conditions, using a monocular camera.

This paper addresses the problem of tracking multiple people in cluttered and overcrowded scenes using a monocular camera. In this scenario the problems encountered include occlusion of persons by other people and background objects, and interactions between subjects. As a consequence, targets can be occluded frequently for long periods of time, making both detection and association challenging. In addition, the problem of tracking is made more difficult by the fact that, viewed from a distance with a relatively low-resolution camera, people tend to have similar appearance, meaning discrimination between people is difficult. There is also the problem of automatically initialising and stopping tracking as people enter and exit the tracked area in an unpredictable manner. All these problems result in highly fragmented tracks and frequent and erroneous identity switches.

Our approach to the problem of multi-target tracking in a realistic environment relies on a robust parts-based pedestrian detector and a dual-stage tracking-by-detection framework to solve the inherent ambiguity of the tracking problem. The first tracking stage uses the output from a robust parts-based pedestrian detector to form short, confident tracks, known as tracklets. In the second tracking stage, the temporal and motion features of each tracklet are learned in order to reliably link tracklets into longer tracks by analysing the information in the gaps between tracklets. The rational behind our dual-stage tracker is to address the problems of track initialisation, occlusion, and long-term tracking. Robust track initialisation is addressed by two features of our framework: Firstly a reliable parts-based pedestrian detector is used to initialise the tracks, and secondly, tracklet initialisation and termination are handled using a hierarchical tracklet confidence scheme, which can cope with the problems of clutter and false-positives. The negative effects of occlusions are reduced by means of an adaptive parts-based appearance model, learned for each person at track initialisation, which is used together with a partial occlusion robust appearance similarity measure, in order to deal with the problem of generating confident tracklets in a crowded and cluttered environment. Finally, long term occlusions are explicitly targeted by the second tracking stage, which links tracklets into longer tracks.

This paper builds on the work in [38] by including a novel tracklet linking method based on motion modelling, which is integrated into a novel hierarchical tracking framework based on the linear assignment problem (LAP). The definitions of the cost functions for both the first and second stages were also improved, in order to obtain improved performance (up to 6% increase in MOTA), a more rigorous formulation of which is also presented. In addition this work includes more extensive experimental evaluation. The rest of the paper is organised as follows: In Section II we describe our proposed pedestrian tracking system. In this section we introduce our novel method of occlusion robust appearance similarity, and describe the second stage of the tracker, that uses our novel method of tracklet linking based on motion modelling using complementary features. Experimental evaluation of the tracking framework is carried out in Section III, and finally in Section IV we present our conclusions.

A. State of the art

Due to the wide range of potential applications, multi-target pedestrian tracking has been extensively studied in the recent past. Early approaches to the multi-target pedestrian tracking problem were based on Kalman filtering [41], [27]. While these approaches have the advantage of simplicity and computational efficiency, making them amenable to real-time implementation, they are prone to identity switches when

there are many closely interacting targets. The recursive nature of such approaches means that it is difficult to detect and correct errors once they have occurred. Alternative ways to optimise the trajectories of objects include the Multiple Hypothesis Tracking algorithm (MHT) [14], [48], which can track a variable number of objects, or the Joint Probabilistic Data Association Filter (JPDAF) [44] which assumes a fixed number of targets. In the limit, MHT builds an exponentially sized tree of all possible target states. However, in practice an approximation such as tree-pruning [48], k-best hypotheses [42] or greedy tracking [51] can be used. It is also possible to use Dynamic Programming to optimise the target trajectories [56], however this method is best suited to tracking a small number of objects, as the computational complexity may become prohibitive when the number of tracked objects grows large.

As opposed to the previously mentioned deterministic tracking solutions, Sequential Monte Carlo methods [31], [36] such as Particle Filtering [10], [43], [13], [20] may be used when Gaussian statistics and linear state models do not apply. Such methods provide a theoretical framework to model and integrate multiple sources of uncertainty considering only the information from past frames. Particle filtering tends to produce continuous tracks, but may diverge from the actual target location. In practice these methods are limited to a few simultaneous targets due to the curse of dimensionality [37] and the difficulty of designing appropriate interaction models [36]. By considering a window of frames, rather than just past frames, the closely related Markov Chain Monte Carlo (MCMC) methods [6] attempt to address the limitations of recursive trackers, while exploring a wider range of tracking hypotheses. Although such methods can achieve high performance, the hypothesis space of a MCMC tracker is exponential in the number of frames considered, leading to increased computational costs. In addition, in order to achieve high performance, these methods may require careful tuning or learning of parameters and ad-hoc interaction models, limiting their applicability in novel situations [8].

Recent advances in pedestrian detectors [15], [17] have made tracking-by-detection approaches practical. In these approaches, prior knowledge that people are the only object-type of interest allows an offline trained pedestrian detector to be used to generate a set of hypotheses for the locations of all the people in each frame. The task of the tracker is reduced to solving the data-association problem, i.e. to group the detections associated with each person into individual tracks [28], [52]. Given a set of pedestrian detections, it is possible to model the problem of the multi-target pedestrian tracking using graph theory. In such an approach the vertices of the graph may either represent discrete world locations, where pedestrians are permitted to exist [7], [18], or they may represent the pedestrian locations hypothesised by a detector [12]. The graph edges are typically used to model the cost of associating two nodes into the same track, based on factors such as appearance similarity [21] or physical distance between detections [12]. Due to the fact that problems in graph theory can be expressed as equivalent linear programs, there are a variety of potential models of the tracking problem including: k-shortest paths [8], flow linear programming [7] and min-cost network flow [12]. While such linear programming methods are appealing due to their mathematically rigorous formulation, certain ad-hoc features such as specialised vertices and edge costs, may be required to represent situations such as missed detections [63] or higher-order motion constraints [33], leading to increased model complexity. Additionally, these methods may require a long sequence of frames before optimisation can be performed, making them unsuitable for low latency applications.

Recently, a variety of methods have been proposed for solving the multi-target pedestrian tracking problem using a hybrid approach, where tracking takes place in two stages. The first tracking stage takes a set of detections output by a pedestrian detector over a short time-window or on a frame-by-frame basis, and produces many short, confident tracks, known as tracklets. These tracklets are very likely to contain only detections from a single person, however they are typically very fragmented. The task of the second stage is to join these tracklets into longer, more stable tracks. The task of producing tracklets, based on linking detections, relies on good detector performance, which may not be available in complex scenarios. One solution to this shortcoming is to incorporate temporal context in order to reduce the impact of false-positives and missed detections [11]. For the task of linking tracklets, a variety of methods have been proposed, ranging from using the Hungarian algorithm, where tracklet association costs are based on the direction of tracklet motion and appearance [53], [22], [52], to association based on parts similarity [28], to using a Conditional Random Field model to take into account higher order costs [58]. In many tracklet linking based approaches [29], [25], the cost of linking tracklets is calculated independently for each tracklet pair, based on factors such as relative velocity and appearance similarity. This approach may lead to globally sub-optimal solutions, if other information is not taken into account [34]. In addition, work has been carried out on tracklet linking approaches that incorporate costs based on social criteria, such as the fact that people tend to walk in groups and to avoid collisions [19], [47], [35]. While trackers incorporating social costs have been shown to improve tracking accuracy, they rely on offline trained social models, therefore their accuracy may be dependent on the similarity between the behaviour of pedestrians in the training scenarios and the observed scenario. Hybrid trackers, based on tracklet generation and linking, do not use a single paradigm to solve the whole tracking problem. Instead, such trackers attempt to compromise between computational efficiency and the use of powerful optimisation methods to solve the difficulties encountered in complex real-world scenarios. The advantage of performing tracking into two stages is that easy tracking decisions can be made by the first stage, significantly reducing the size of the hypothesis space. Then, the second stage, using more sophisticated reasoning can be used to solve the tracking problem in complex scenarios.

In this work we propose a novel tracker that uses a two-stage hybrid approach, to solving the pedestrian tracking problem in realistic scenarios. Our contribution is two-fold. Firstly, in

our first stage, we propose the use of a novel occlusion robust similarity measure for linking detections into tracklets. This is important as partial occlusion is a frequent occurrence in crowded scenarios. Secondly, in our next stage, we propose the use of a novel tracklet linking process, based on optical flow patterns in the gaps between tracklets. By using this additional information, not contained in the tracklets themselves, we hope to avoid the scenario where tracklets are linked in a sub-optimal way, based only on local features such as velocity and appearance.

B. Tracking using the Linear Assignment Problem

Assuming the existence of a set of detections generated by a pedestrian detector, the tracking problem can be solved in two stages: firstly detections are joined into short confident tracks, known as tracklets, and secondly, tracklets are linked together into longer tracks. Both stages can be modelled as finding the optimal solution to a linear assignment problem (LAP) [45], where each detection can only be assigned to a single tracklet, and tracklet linking must be pairwise.

For the first stage, assuming an equal number of tracklets and detections, the LAP problem can be modelled using two equal-size square matrices: cost matrix C and assignment matrix A, where $c_{i,j} \in C$ denotes the cost of assigning detection i to tracklet j, and entry $a_{i,j} \in A$ indicates that detection i has been assigned to tracklet j if $a_{i,j} = 1$. The optimal solution to this assignment problem can be defined as

$$\hat{A} = \arg\min_{A} \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{a}_{i,j} c_{i,j} \tag{1}$$

where N is the size of the matrices, and the optimal solution $\hat{A}$ is subject to the constraints

$$\sum_{l=1}^{N} \hat{a}_{l,k} = 1, \quad \sum_{k=1}^{N} \hat{a}_{l,k} = 1 \quad \forall l, k \in [1 \dots N] \tag{2}$$

The optimal assignment matrix $\hat{A}$, which minimises the global cost, can be found in polynomial time using the Hungarian algorithm [32]. In reality, false-positives and missed detections mean there are unequal numbers of detections and tracklets. In this case, an augmented cost matrix $\hat{C}$, can be formed, where the upper left corner contains rectangular cost matrix C, the upper right and lower left corners contain diagonal matrices, containing costs for unassociated detections, and the lower right corner of $\hat{C}$ need not contain any values. The costs for unassociated detections should be higher than any other costs in the original C matrix, to prevent non-associations unless there is no alternative solution.

In the second tracker stage, LAP can again be used to model linking of tracklets into longer tracks within a time-window [29], using a new tracklet linking cost matrix C, where $c_{i,j}$ may depend on factors such as motion smoothness, time difference, and appearance similarity.

II. Pedestrian Tracking Framework

We propose a novel tracking system based on the LAP framework, that is adapted to the problem of pedestrian tracking in realistic scenarios. The original formulation of the LAP tracking framework [29] approximates the data association component of MHT by independently linking detections between each pair of frames. We have adapted this method for online pedestrian tracking by introducing the concept of a sliding window. We also propose to add a prediction/estimation filter in order to adapt the first stage of the LAP framework to the problem of pedestrian tracking, based on the observation that people move in a predictable manner over short periods of time. In addition, we will use a hierarchical tracking architecture, where easy tracking cases are solved using a greedy approach, to produce short reliable tracklets, which are then linked into longer confident tracks using optimisation over a window of frames. Within this overall tracking framework we introduce two main novel contributions:

In the first stage, where detections are linked to tracklets, we propose a novel occlusion robust appearance similarity measure, employed in the cost-matrix for linking tracklets with detections at every frame. This method improves the quality of tracklets produced under realistic crowded conditions, where partial occlusions may frequently occur, compared to conventional methods of finding the appearance similarity between tracks and detections, as will be shown in Section III.

In the second stage, short tracklets are linked together into longer tracks. Here we propose to use a novel gating function for determining which tracklets may be linked, on the basis of an online classifier that uses optical flow features to validate proposed links between tracklets. Optical flow models are learned online for each tracklet, as well as the scene background, and only those links consistent with the optical flow evidence are permitted. This approach removes the need for ad-hoc rules governing the linking of tracklets, based on factors such as relative angle or motion smoothness.

A. Overview of Tracking Framework

A block diagram of our tracking approach is depicted in Fig. 1. There are three main parts: detection and filtering, tracklet generation and tracklet linking. We initially apply a pedestrian detector to each frame, the detections are then filtered using height and non-maxima suppression to reduce the number of false positives. In the first tracker stage, detections are associated with corresponding tracks, taking into account occlusion using our novel occlusion robust appearance similarity method. Data-association is performed hierarchically, taking into consideration the evidence associated with each tracklet. To address more complex scenarios, where the first stage tracker may be insufficient, a second tracking stage is introduced. In an ongoing process, tracklet confidence is assessed by the first stage, and confident tracks are passed to the second tracker stage where optical flow appearance models are learned for tracklets and the gaps between tracklets. These optical flow models are then used to reliably link fragmented

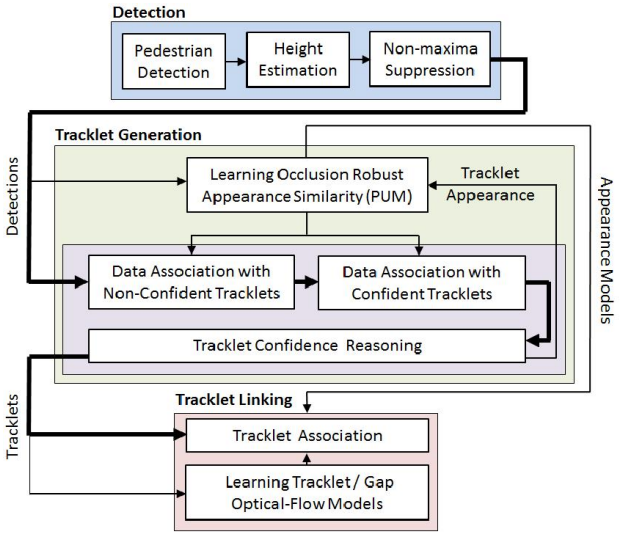


Fig. 1. Flowchart of our tracking system.

tracklets into longer tracks, in order to cope with the limited reasoning capabilities of the first stage, where linking between tracklets and detections is performed on a frame-by-frame basis.

B. Tracklet Generation

The tracklet generation mechanism used in the first stage is a greedy online tracker based on the Markovian assumption that the state at time t depends only on the state at time $t-1$ and the observation at time t. At each time-step this online Markovian tracker receives a new set of detections, each of which is either associated with an existing tracklet or used to initialise a new tracklet. Similarly to the approach used in [52], the Markovian tracker uses a tracklet hierarchy consisting of confident and non-confident tracklets. Confident tracklets have been associated with many detections and passed several initialisation requirements, while non-confident tracklets have only been associated with a smaller number of detections and have not passed the initialisation requirements to be considered confident tracklets. All tracklets are initialised as non-confident tracklets. A non-confident tracklet may be promoted to the status of confident tracklet upon meeting the following initialisation requirements: Firstly, within a time window of duration β frames, the number of detections associated with the non-confident tracklet must be greater than α. And secondly, the speed of the tracklet during this window must be greater than zero, in order to rule out false positives generated by static background objects.

Tracklets are terminated if they have not been associated with a detection for a time-period of υ frames, causing tracklet confidence to drop, or if they have reached the edge of the image, where they are assumed to be exiting the scene.

To cope with short-term occlusions and temporary detector failure, tracklets are allowed to drift for a short time-period without being associated to a detection. During this time-period the tracklet position is predicted using the state model of its associated Kalman filter. At each time-step the Kalman filter predicts a smoothed estimate of the tracklet's state using the observation at time t and the state at time $t-1$. The Kalman filter is used to better estimate the state of each tracklet at each time-step, given the set of noisy detections previously associated with the tracklet. It allows the inclusion of the concept of prediction, which is absent in the LAP formulation of [29]. The state vector $\mathbf{s} = (p_x, p_y, \dot{p}_x, \dot{p}_y, s_x, s_y, \dot{s}_x, \dot{s}_y)$, of each tracklet in the Kalman filter consists of its world-position, velocity, bounding-box size and rate-of-change of bounding-box size. A linear motion model was assumed for both position and bounding box size.

Data association between new detections and the predicted state of existing tracklets is performed hierarchically. First, association is attempted between new detections and confident tracklets. Association is then attempted between any remaining unassociated detections and non-confident tracklets. Finally, any still remaining detections are used to initialise new non-confident tracklets. This hierarchical approach to data association between tracklets and detections is important for the generation of reliable tracklets [60], [61], [52]. In a system with no distinction between confident and non-confident tracklets, it would be possible for a newly created false positive tracklet to 'steal' detections from an established tracklet. A hierarchical data-association approach based on confidence prevents such situations, as association decisions are made in a way that respects the weight of supporting evidence associated with each class of tracklet.

When performing data-association, the cost of linking tracklet $\mathbf{t}_i$ with detection $\mathbf{d}_j$ is defined as

$$C(\mathbf{t}_i, \mathbf{d}_j) = \frac{1}{P(\mathbf{t}_i|\mathbf{d}_j)B(\mathbf{t}_i, \mathbf{d}_j)} \tag{3}$$

where $P(\mathbf{t}_i|\mathbf{d}_j)$ is the appearance similarity of tracklet $\mathbf{t}_i$ given detection $\mathbf{d}_j$, and $B(\mathbf{t}_i, \mathbf{d}_j)$ is the degree of overlap between the bounding boxes of detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$. Calculation of the appearance similarity $P(\mathbf{t}_i|\mathbf{d}_j)$, which takes into account the possibility of partial occlusion, is discussed in Section II-C. The value of the appearance similarity is bounded in the range 0 to 1, reaching its maximum value when the appearances of the tracklet and detection are identical. The degree of overlap $B(\mathbf{t}_i, \mathbf{d}_j)$ between the bounding boxes of detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$ is defined as

$$overlap = \frac{Area(\mathbf{t}_i \cap \mathbf{d}_j)}{Area(\mathbf{t}_i \cup \mathbf{d}_j)} \tag{4}$$

$$B(\mathbf{d}_i, \mathbf{t}_j) = \begin{cases} overlap & overlap \geq \tau \\ 0 & overlap < \tau \end{cases} \tag{5}$$

where the threshold τ in the gating function is used to ensure that association can only take place between tracklets and detections with a large degree of overlap, thus preventing many potentially incorrect associations. The overlap function is zero when there is no overlap of the bounding boxes, and reaches a maximum value of one when the bounding boxes are identical in size and position.

Association between tracklets and detections is modelled as a LAP where each detection may only be associated with a

single tracklet, based on the simplifying assumption that each true pedestrian will generate a single detection in each frame. The minimum cost solution to this assignment problem can be efficiently computed using the Hungarian algorithm [32]. An assignment matrix between tracklets and detections is created, where the cost of assigning detection $\mathbf{d}_i$ to tracklet $\mathbf{t}_j$ is

$$A(\mathbf{d}_i, \mathbf{t}_j) = \begin{cases} C(\mathbf{t}_i, \mathbf{d}_j) & B(\mathbf{d}_i, \mathbf{t}_j) > 0 \\ \infty & B(\mathbf{d}_i, \mathbf{t}_j) = 0 \end{cases} \quad (6)$$

where $C(\mathbf{t}_i, \mathbf{d}_j)$ is the cost of linking detection $\mathbf{d}_i$ with tracklet $\mathbf{t}_j$ as defined in Eq. (3), and where $B(\mathbf{t}_i, \mathbf{d}_j)$ is a gated function of the degree of overlap between the predicted location of tracklet $\mathbf{t}_i$ and the detection $\mathbf{d}_j$ as defined in Eq. 4. By limiting associations to tracklets with significantly overlapping detections, false associations due to detector failure and coincidental appearance similarity are reduced, resulting in more reliable tracklets.

C. Occlusion-Robust Appearance Similarity

Appearance is one of the most discriminative features that can be used to resolve tracking ambiguity when spatio-temporal features are not sufficient, such as can happen in overcrowded scenes. However, data-association between tracklets and detections based on appearance can be difficult due to the interaction of people with background objects and other pedestrians, which can lead to occlusion.

In order to take the possibility of occlusion into account, we incorporate a novel method to compute an occlusion robust appearance similarity between tracklets and detections. This method is based on the posterior union model (PUM) [40]. This missing feature method has previously been used for speaker identification in noisy conditions [40], and facial recognition given partial occlusion [39]. Missing feature methods are used to calculate a matching score, between a training example and a partially corrupted test sample, while ignoring the contribution of the corrupted parts to the overall score. The advantage of the PUM is that such a score can be calculated without the need for explicit knowledge of which parts are corrupted. This makes the PUM useful for tracking applications where it can be difficult to accurately determine the occluded regions.

Assume that the appearances of the detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$ can be represented by sets of corresponding parts. Let $\mathbf{t}_i = (t_i^1, t_i^2, \ldots, t_i^n)$ represent the appearance model for tracklet $\mathbf{t}_i$, composed of n appearance features extracted from the major body parts, and let $\mathbf{d}_j = (d_j^1, d_j^2, \ldots, d_j^n)$ represent the corresponding appearance model of detection $\mathbf{d}_j$. In a conventional approach, the appearances of corresponding parts would be compared and the matching scores then combined, e.g. by summation or multiplication, to produce a score for the overall appearance similarity between detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$. Given partial occlusion, some of the matching scores will be corrupted. We can express the appearance similarity $P(\mathbf{t}_i|\mathbf{d}_j)$ between detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$, taking into account partial occlusion as

$$P(\mathbf{t_i}, |\mathbf{d}_j) \propto \max_{X_s} P(\mathbf{t}_i|\mathbf{d}_j, X_s) \quad (7)$$

where the index set $X_s \subseteq [1 \ldots n]$, is used to denote a subset of the appearance features of m elements, where $m \leq n$, and $P(\mathbf{t}_i|\mathbf{d}_j, X_s)$ expresses the posterior probability of tracklet $\mathbf{t}_i$ given detection $\mathbf{d}_j$ and the optimal index set of matching features X_s. Using Bayes formula $P(\mathbf{t}_i|\mathbf{d}_j, X_s)$ can be expressed as follows

$$P(\mathbf{t}_i|\mathbf{d}_j, X_s) = \frac{P(\mathbf{d}_j|\mathbf{t}_i, X_s)P(\mathbf{t}_i)}{\sum_{t_k} P(\mathbf{d}_j|\mathbf{t}_k, X_s)P(\mathbf{t}_k)} \quad (8)$$

where the denominator sums over the posterior probability of all tracklets $\mathbf{t}_k$. Note that we assume each tracklet has equal prior probability, meaning the prior term is redundant and can be dropped from the calculation. One problem with this approach is that given a set with N elements, there are 2^N possible subsets of features given an unknown number of m. Therefore for even moderately sized sets, the search for the optimal feature subset is very inefficient. Following the approach in [40] we can reduce the complexity of this search by approximating $P(\mathbf{d}_j|\mathbf{t}_i, X_s)$ with the union of all feature subsets of equal size i.e. $P(\mathbf{d}_j|\mathbf{t}_i, X_s) \propto P(\mathbf{d}_j|\mathbf{t}_i, X_{s(m)})$ where $X_{s(m)}$ is the union of all feature subsets of size m. This approximation can be efficiently computed using dynamic programming [40], reducing the complexity of finding the optimal subset from $O(2^N)$ to $O(N^2)$. We can therefore define an approximation of the posterior probability $P(\mathbf{d}_j|\mathbf{t}_i, X_s)$ of detection $\mathbf{d}_j$ given tracklet $\mathbf{t}_i$ and the index set X_s as

$$P(\mathbf{d}_j|\mathbf{t}_i, X_{s(m)}) \equiv \sum_{|X_s|=m} \prod_{l \in X_s} M^{b(\mathbf{t}_i^l, \mathbf{d}_j^l)} \quad (9)$$

where M is a positive base number. As we represent the appearance of each part using a colour histogram, $b(\mathbf{t}_i^l, \mathbf{d}_j^l)$ is the Bhattacharyya coefficient between the histograms representing appearance features $\mathbf{t}_i^l$ and $\mathbf{d}_j^l$. The value of M used should be large, so that the optimal feature subset will dominate the summation i.e. its value will be much greater than any other feature subset with the same cardinality. Finally, the occlusion robust similarity $P(\mathbf{t}_i|\mathbf{d}_j)$ between detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$ can be defined as

$$X_s \propto \arg\max_m \frac{P(\mathbf{d}_j|\mathbf{t}_i, X_{s(m)})}{\sum_{t_k} P(\mathbf{d}_j|\mathbf{t}_k, X_{s(m)})} \quad (10)$$

$$P(\mathbf{t}_i|\mathbf{d}_j) \propto \max_{X_s} \frac{P(\mathbf{d}_j|\mathbf{t}_i, X_{s(m)})}{\sum_{t_k} P(\mathbf{d}_j|\mathbf{t}_k, X_{s(m)})} \quad (11)$$

where maximisation over the feature subset size m is carried out to find the optimal subset of matching features. Note that as mentioned previously, the prior term has been dropped due to the assumption that all tracklets have equal prior probability.

D. Unsupervised Tracklet Linking by Gap Modelling

The first stage of the tracker produces a set of short confident tracklets, which are linked by the second stage to form long confident tracks. This tracker architecture avoids the need to make each tracking stage overly complex. The first tracking stage is based on greedy data-association of tracks with detections, solved by the Hungarian algorithm. This process is optimal at each time-step, but may not produce

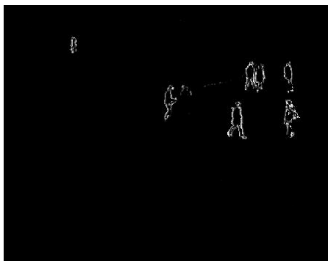


Fig. 2. Magnitude optical flow image from the PETS 2009 S2.L1. Due to the static camera and elevated positioning, pedestrians are clearly distinguishable from the background. (Contrast enhanced for display purposes).

the global optimum assignment of detections into tracklets, as the decisions cannot be later revised. Hence, only links with very high certainty are accepted, in order to reliably solve easy linking cases. At some point, the features used for linking detections with tracklets in the first stage are no longer sufficient, and the second stage tracklet linking process must take over. The second tracking stage must be capable of solving more difficult cases such as long term detector failure or larger appearance changes. The second stage solves these problems by reasoning over a window of frames, using additional information compared to the first stage tracker.

Many existing trackers solve the tracklet linking problem using ad-hoc rules governing which tracklets may be linked together. These kinds of tracklet linking approaches may not be scalable to more complex scenarios as the parameters of the linking model must be learned or adapted to novel scenarios [24]. We therefore propose a novel tracklet linking method based on modelling tracklets, and the space between tracklets, or gaps, using optical flow features. This gives a natural method for deciding if tracklets should be linked, and avoids the need for ad-hoc rules. The optical flow features are used to build a gating function for the tracklet linking process, so that only links that are consistent with the observed optical flow evidence are permitted.

We can think of a video as a space-time volume, where optical flow provides information about the direction and magnitude of motion occurring at each space-time position. People moving in video will leave a continuous optical flow signature in the volume that can be used to help constrain the tracklet linking problem. An example of an optical flow image taken from a static surveillance camera is shown in Fig. 2, note that pedestrians are clearly visible and their associated optical flow is very different from the background optical flow. Background optical flow refers to any optical flow signal not caused by the motion of a person, for example, optical flow measurements from other types of moving objects, or spurious readings caused by noise. The appearance of the background optical flow is quite different to that of a moving pedestrian.

Optical flow has previously demonstrated its potential for detecting and tracking people when using conventional methods becomes difficult due to crowded conditions [2], [26], [49]. In our novel approach, we propose to model the optical flow signature of each person, allowing the tracker to utilise information about the velocity and unique motion pattern associated with each person [55], [50] to calculate the likelihood of all proposed links originating at the tracklet, given the observed optical flow evidence. This optical flow signature can be used in conjunction with a model of the background motion optical flow, or gap model, to create a gating function for the tracklet linking process. Links are only permitted if they are consistent with the observed optical flow evidence, removing the need for ad-hoc rules governing the tracklet linking process and simplifying the tracklet linking cost function. This approach has the advantage that it can generalise to many different tracklet linking scenarios, without the need for a large number of parameters e.g. describing the permissible range of tracklet linkages based on motion smoothness.

The optical flow model of each tracklet, and the background motion model, are learned online as a histograms of features. For each tracklet, the model is averaged over the whole length of the tracklet. The model for the source tracklet $\mathbf{t}_s$ can be defined as a histogram $h(\mathbf{t}_s) = \{\mathbf{t}_s^m\}_{m=1..M}$ with M bins, where $\mathbf{t}_s^u = \frac{1}{Z}\sum_{t=1}^{T}\sum_{i=1}^{I}\delta\left[of(x_{i,t}) - u\right]$, T is tracklet length in frames, I is the number of optical flow voxels in the tracklet bounding box, $of(x_{i,t})$ is the magnitude and phase value i at time t, δ is the Kronecker delta function, u is the centre value of the histogram bin, and Z is a normalisation factor ensuring that the histogram sums to one. The background motion model can be defined in a similar manner as, $h(bg) = \{bg^m\}_{m=1..M}$ where $\mathbf{bg}_s^u = \frac{1}{Z}\sum_{k=1}^{K}\sum_{i=1}^{I}\delta\left[of(x_{i,k}) - u\right]$. This background motion model is built using K random samples taken from the optical flow image, with care taken to ensure that these samples do not overlap $> 50\%$ with the bounding box of any tracklet.

Optical flow evidence is integrated into the tracker framework as a gating function during tracklet association, to disallow links that are inconsistent with the optical flow evidence. To decide if a link between tracklets is consistent with the optical flow evidence the following steps are performed: For each pair of tracklets with a feasible link i.e. where the end of the source tracklet $\mathbf{t}_s$ and the beginning of the target tracklet $\mathbf{t}_t$ fall within the currently considered time-window, the path linking the tracklets is interpolated and smoothed. At each time-step t along the proposed path, the optical flow evidence $\mathbf{o}_{s,t}(t)$ is compared with the optical flow model of the source tracklet $h(\mathbf{t}_s)$, the background motion model $h(bg)$, and the models for all other tracklets $h(\mathbf{t'}_k)$. The gating function $L(\mathbf{t}_s, \mathbf{t}_t)$ is calculated using the likelihood ratio of the optical flow evidence given the source tracklet model, compared to the optical flow models for all other tracklets, and the background motion model:

$$L(\mathbf{t}_s, \mathbf{t}_t) = \frac{1}{T}\sum_{t=1}^{T}\frac{P(\mathbf{o}_{s,t}(t)|h(\mathbf{t}_s))}{P(\mathbf{o}_{s,t}(t)|h(bg)) + P(\mathbf{o}_{s,t}(t)|h(\mathbf{t'}_k))} \tag{12}$$

where T is the period of time, measured in frames, from the beginning of the source tracklet $\mathbf{t}_s$ to the start of the target tracklet $\mathbf{t}_t$, and $P(\mathbf{o}_{s,t}(t)|h(\mathbf{t'}_k)$ is the average likelihood of the optical flow evidence given all other tracklet models. The likelihoods $P(\mathbf{o}_{s,t}(t)|h(\mathbf{t}_s))$ and $P(\mathbf{o}_{s,t}(t)|h(bg))$

are calculated using the Bhattacharyya coefficient between the histograms of the optical flow image evidence and the optical flow models for the source tracklet $\mathbf{t}_s$ and the optical flow background model $h(bg)$. Similarly, $P(\mathbf{o}_{s,t}(t)|h(\mathbf{t'}_k))$ is calculated as the average Bhattacharyya coefficient between the observed optical flow, and the optical flow models of all other tracklets $\mathbf{t'}_k$.

Finally, association between tracklets is modelled as a LAP within each time-window, or length ζ frames. For each pair of tracklets a cost function is evaluated to determine the likelihood of a link between the tracklets. The cost of linking a pair of tracklets is defined as

$$A(\mathbf{t}_s, \mathbf{t}_t) = \begin{cases} \frac{1}{P(\mathbf{t}_s|\mathbf{t}_t)} + \Delta D(\mathbf{t}_s, \mathbf{t}_t) + \Delta T(\mathbf{t}_s, \mathbf{t}_t) & L(\mathbf{t}_s, \mathbf{t}_t) \geq 1 \\ \infty & L(\mathbf{t}_s, \mathbf{t}_t) < 1 \end{cases} \quad (13)$$

where $P(\mathbf{t}_s|\mathbf{t}_t)$ is the occlusion robust appearance similarity defined in Section II-C which is used here to calculate the appearance similarity between tracklets, and where $\Delta D(\mathbf{t}_s, \mathbf{t}_t)$ and $\Delta T(\mathbf{t}_s, \mathbf{t}_t)$ are monotonically increasing functions of the Euclidean distance, and time difference, between the tracklets respectively. The tracklet linking problem is, therefore, first simplified using the gating function based on motion modelling $L(\mathbf{t}_s, \mathbf{t}_t)$, and then solved based on visible appearance, distance and time. Using the above cost function, the minimum cost association between tracklets within the current time window is found using the Hungarian algorithm.

III. Experimental Evaluation

The tracker was evaluated on two standard public datasets, the town centre dataset [6] and the PETS 2009 dataset [1]. The town centre dataset contains a realistic street scenario captured with a resolution of 1920×1080 pixels at 25 fps, for 4500 frames. It features naturalistic pedestrian behaviour, with many cases of short-term partial and full occlusion, as well as several cases of long-term occlusion. The crowd density varies from sparse to moderately crowded, with an average of 16 people in each frame, over the whole sequence. The PETS 2009 dataset features several staged scenarios of pedestrians walking in a university campus, captured with a resolution of 768×576 pixels at 7 fps. The PETS S2.L1, S2.L2 and S2.L3 sequences were used to facilitate comparison with the literature. The crowd density varies from sparse in S2.L1 (average of 6 people per frame), to densely crowded in S2.L3. These sequences contain many cases of partial occlusion caused by other pedestrians, and full occlusion caused by a static foreground object. In places, the motion of pedestrians is unpredictable, with several cases of abrupt reversal of walking direction. The Oxford and PETS sequences have been used extensively in recent tracking papers, facilitating comparison with the state of the art.

The track-initialisation parameters (see Section II-B) were set as follows. The parameter β, which specifies the window length used for track-initialisation, was set to a number of frames equivalent to 1 second. In the case of the town centre sequence this was 25 frames, while for PETS sequence this was 7 frames. The parameters α, which specifies the detection-rate i.e. the number of detections within the window β a non-confident tracklet must accumulate before being considered confident, must be adapted depending on the detector reliability. For the Oxford dataset $\alpha = 20$, for PETS S2.L1, with low crowd density $\alpha = 6$, for S2.L2 with moderate density $\alpha = 3$, and for S2.L3 with high crowd density $\alpha = 2$. The parameter υ, specifying the time-period a track is allowed to drift without receiving a detection, was 2α, and ζ, the parameter specifying the window length for tracklet linking, was set to the number of frames equivalent to 2 seconds. The Kalman filter parameters, were set to values that produced good results across a range of test sequences.

Fig. 3. Body parts used for modelling appearance. Left - Pedestrian bounding box showing parts corresponding to the head, body, and upper legs, where colour histogram features are extracted. Right - A frame from the Oxford sequence with corresponding body parts displayed.

Pedestrians were detected using the Poselets detector [9]. Similar reasoning to [10] was followed to retain the true positive detections, by considering as valid detections even ones with low confidence. False positives were filtered by using the camera calibration to remove detections outside the normal range range of human height [54]. A non-maxima suppression (NMS) algorithm [16] also reduces the number of false positives surrounding confident pedestrian detections.

The appearance of each detection $\mathbf{d}_j$ and tracklet $\mathbf{t}_i$ (see Section II-C) was modelled using 3D colour-histogram features extracted from the main body parts, localised by the Poselets detector, as seen in Fig. 3. For example, given a detection $\mathbf{d}_j$, the appearances of parts $d_j^1 \ldots d_j^n$ were represented using 3D colour-histograms extracted from the head, torso and upper legs regions respectively, where $n = 3$ parts were used. The lower legs were not included due to their small size and large motion, which resulted in background information being included in the appearance model.

To cope with appearance variation over time, each tracklet's appearance model is updated at every frame using η histograms retained from previously associated detections. In our implementation η is set to a number of frames equivalent to 3 seconds. The appearance model is calculated as the median of all the retained features, allowing the model to cope with short-term occlusions and temporarily incorrect associations, while remaining adaptive to long-term appearance variation.

For evaluation we use the standard CLEAR MOT performance metrics [30], and tracks are associated with the ground-truth using the standard PASCAL 50% overlap criterion [6]. The CLEAR MOT metric defines two measures of tracker

performance: multiple object tracker accuracy (MOTA), and multiple object tracker precision (MOTP), as follows:

$$MOTA = 1 - \frac{\sum_t^T m_t + fp_t + mme_t}{\sum_t^T g_t} \quad (14)$$

$$MOTP = \frac{\sum_{i,t} d_i^t}{\sum_t c_t} \quad (15)$$

where in the definition for MOTA, for a given frame t in a sequence of length T, we define, m_t the number of false-negatives (missed-detections), fp_t the number of false-positives, mme_t the number of ID-switches, and g_t the number of ground-truth objects. In the definition of MOTP we define d_i^t as the distance between the location of ground-truth object i and the associated detection, and c_t as the number of ground-truth objects in each frame. MOTA is widely accepted as a good reflection of true tracker performance, as it measures false-positives, false-negatives and ID-switches, whereas MOTP simply measures how closely the tracker follows the ground-truth, regardless of any other errors. For this reason MOTA will be occasionally used stand alone to evaluate some of the tracking tuning decisions. We also evaluate tracking performance using the average precision $\frac{1}{T}\sum_t^T \frac{tp_t}{tp_t+fp_t}$ and average recall $\frac{1}{T}\sum_t^T \frac{tp_t}{tp_t+fn_t}$, where tp_t, fp_t and fn_t, are defined as the number of true-positives, false-positives and false-negatives at frame t respectively.

A. Tracklet Generation

In the first tracking stage, two factors govern data association between tracklets and detections: appearance similarity, and the overlap between the predicted tracklet location and detection bounding box. In this experiment, the effect of both factors on tracking performance is investigated.

Firstly, our proposed occlusion robust appearance similarity (See Section II-C) is compared with alternative methods for combining the appearance scores of individual body parts: sum of all parts similarities, product of all parts similarities, and a holistic appearance model using a single 3D colour histogram to represent the whole bounding box.

Secondly, the bounding box overlap threshold τ (see Eq. 4), which gates association between tracklets and detections, is varied from zero to one. Progressively increasing τ increases the strictness of match required for data association, as association is only permitted if the overlap between the predicted tracklet location and detection bounding box is greater than τ. This experiment was carried out using all 4500 frames of the Oxford town centre dataset.

It can be seen from the results in Fig. 4 that the overlap threshold τ has a large effect on overall tracker accuracy. Initially, increasing τ increases tracker accuracy, for all appearance models, until a certain point, then tracker performance drops sharply. At first, increasing τ reduces the number of false positive detections included in the data association step, while retaining correct detections, thus increasing tracker performance. Performance drops when the value of τ becomes very high, as many correct detections are then rejected. Maximum accuracy occurs when $\tau = 0.55$, but to prevent over-fitting to a specific dataset, all experiments in the following sections use $\tau = 0.5$.

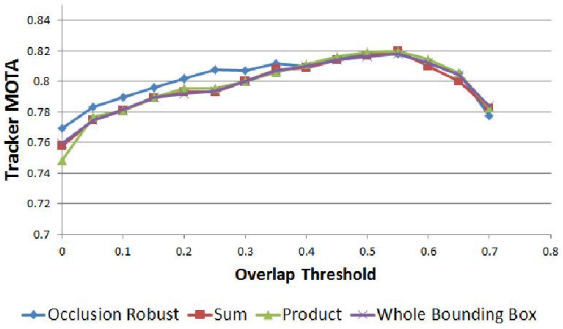


Fig. 4. Tracker performance, measured using MOTA, varies as a function of both the appearance similarity method used, and the overlap threshold τ used in the gating function for association between tracklets and detections. Note that we only show the results for τ in the range 0 to 0.7, as the MOTA sharply declines after this point.

Regarding the evaluation of the PUM contribution, the results in Fig. 4 show that the occlusion robust appearance similarity was better able to correctly associate tracklets with detections at low overlap thresholds, where it out-performed all other appearance similarity methods that did not explicitly consider occlusion. As the overlap threshold was increased to near 0.5, all the appearance similarity methods started to behave identically, as at this point there are very few distractors competing with the correct detection for each tracklet, and appearance similarity ceases to play a major role in the association process. Thus it can be concluded that PUM provides the best performance by considering occlusions affecting the appearance similarity, and that this increase of performance is more noticeable given more complex association scenarios. This is particularly important for realistic video surveillance sequences with low resolution and low frame-rates, where it is not possible to precisely tune the data association parameters, and where there may be deterioration in detector performance leading to more complex data association scenarios.

B. Parameter Sensitivity

The parameter α, which specifies the detection-rate a tracklet must achieve during initialisation to be considered a confident tracklet (see Section III). The optimal value for this parameter is related to the detector reliability, which in turn depends on factors such as the crowd density in a particular sequence. To test the sensitivity of our tracking system to this parameter, we measured tracker performance over a wide range of α values, for the PETS and Oxford sequences.

The results in Fig. 5 show that the performance of our system is relatively stable over a wide range of α values. In order to achieve optimal performance, the value of this parameter can be adjusted to reflect the specifics of a particular sequence. For instance, in the PETS S2.L1 sequence, with low crowd density, best performance is achieved when α takes a high

value. In contrast, the S2.L2 and S2.L3 sequences have high crowd density, and hence low detector reliability, therefore best performance is achieved when α takes a low value. The Oxford sequence has moderate to sparse crowd density, and good performance is observed over a wide range α values. However, in all cases an intermediate value of α will give good performance in all conditions, as illustrated in Table. I.

C. Tracklet Linking Using Optical Flow

In this experiment we investigate how the use of optical flow based tracklet linking can improve tracking performance by comparing: optical flow gated tracklet linking, tracklet linking without optical flow gating, and the tracker used with no tracklet linking. The cost function of the no optical flow linking method was identical to the optical flow method, Eq. (13), except that the optical flow gating function was omitted (see Section II-D). Instead links were permitted between all tracklets within the current time-window. In this experiment, for all PETS sequences the parameter α was held constant at $\alpha = 4$, as the goal was to investigate the relative performance of the tracklet linking methods, rather than to evaluate the best potential performance. The results of this experiment are shown in Table I.

Sequence	Method	MOTA	MOTP	Prec.	Recall
PETS S2.L1	OF Linking	79.56	68.34	88.89	90.92
	No OF Linking	78.26	68.36	89.53	88.62
	1^{st} Stage Only	77.85	68.35	89.49	88.21
PETS S2.L2	OF Linking	54.43	71.80	89.17	61.96
	No OF Linking	53.56	71.97	91.60	58.97
	1^{st} Stage Only	53.33	71.98	91.89	58.50
PETS S2.L3	OF Linking	51.36	73.62	91.50	56.62
	No OF Linking	50.38	73.81	93.62	54.06
	1^{st} Stage Only	50.17	73.82	93.99	53.60
Oxford	OF Linking	80.78	70.85	95.33	85.24
	No OF Linking	80.68	70.85	95.38	85.08
	1^{st} Stage Only	80.67	70.85	95.39	85.06

TABLE I
COMPARISON OF TRACKER PERFORMANCE USING DIFFERENT TRACKLET LINKING METHODS: OPTICAL FLOW BASED TRACKLET LINKING - OF LINKING. TRACKLET LINKING WITH NO OPTICAL FLOW LINKING - NO OF LINKING. NO TRACKLET LINKING - 1^{st} STAGE ONLY.

From the results in Table I it can be seen that, in terms of MOTA, the optical flow based tracklet linking method

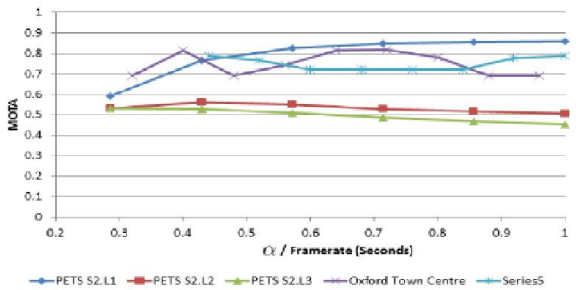
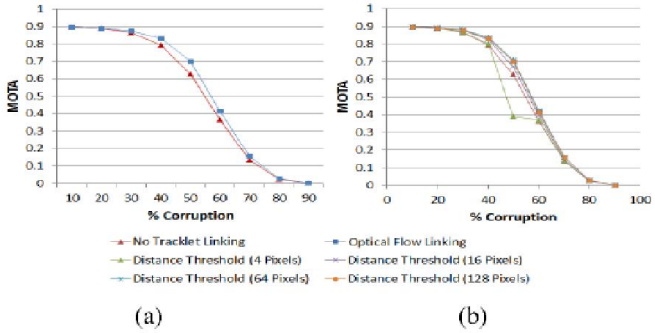


Fig. 5. Tracker performance as parameter α, specifying the detection-rate used during tracklet initialisation, is varied. Note that α is displayed in terms of detections per second.



(a) (b)

Fig. 6. Tracklet linking with simulated tracklets. (a) Compares the performance of No Tracklet Linking and Optical Flow Linking. (b) Compares tracklet linking using a distance threshold based gating function, with various thresholds, to Optical Flow Linking and No Tracklet Linking.

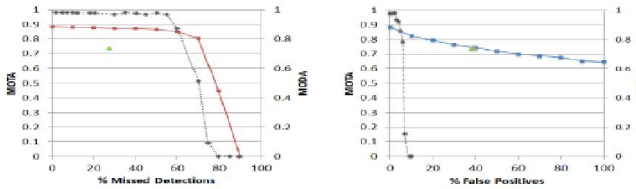
consistently improves performance, compared to the system tested with no tracklet linking and tracklet linking without optical flow gating. When optical flow tracklet linking is used, recall improves significantly as tracklets are correctly joined, thus filling gaps in the tracks, while precision falls slightly as some tracklets are incorrectly joined. For the Oxford sequence, use of optical flow tracklet linking gives only a small improvement, as performance is already very high, meaning there are fewer fragmented tracklets available to potentially join.

In a second experiment, to demonstrate tracklet linking performance in a systematic manner, tracklets were generated using the Oxford ground-truth as a simulated detector, with a percentage of detections randomly dropped at each frame. Additionally, the first stage was modified to produce more fragmented tracklets. The results in Fig. 6 (a) show that our optical flow tracklet linking method is most effective given a moderate to high degree of corruption. At low corruption, with little fragmentation, tracklet linking cannot make a significant difference, while at high corruption, there are few reliable tracklets to be linked.

In the third part of this experiment, we perform the same experiment to compare the performance of our tracklet linking against a baseline tracklet linking system based on distance and trajectory smoothness, such as [22]. Links are permitted if the distance between the tracklets is less than a threshold, and the angle is smaller than 45 degrees. The results in Fig. 6 (b) show that, while it is possible to tune the distance-based gating parameter to give results comparable with the optical flow gating method, incorrectly tuning this parameter detrimentally affects tracker performance. In contrast, our linking strategy based on an optical flow gating function produces improved results, without the need to tune any parameters, and can automatically adapt to varying conditions, producing improved results, compared to the first stage alone, in several different sequences, using both real, and simulated tracklets.

D. Sensitivity To Pedestrian Detector Performance

To understand the role of detector performance in tracking, we used the ground-truth detections from the Oxford sequence



(a) (b)

Fig. 7. (a) The solid red line shows how tracker performance (MOTA) is affected by increasing the number of missed detections. (b) The solid blue line shows how tracker performance (MOTA) is affected by simulating an increased number of false positives. The dotted grey line shows the performance (MODA) from a similar experiment in [8], and the green triangle shows tracker performance (MOTA) given know detector accuracy from [3].

as a benchmark for tracker performance assuming a near-perfect detector. We then proceeded to independently vary the number of false positives and missed detections. Note that, this is different from simply varying the confidence threshold of an existing detector, as that would vary both types of error together.

In the first part of this experiment a random percentage of detections, from 0% to 90%, was dropped at each frame, while all other tracker parameters remained constant. The results in Fig. 7 (a), show that the tracker is not highly sensitive to missed detections. Performance remains approximately constant until around 60% of detections have been dropped. Below this level, tracker performance remained near 0.9, which is still much higher than when using the Poselets detector (See Section III-G), due to the absence of observation noise in the simulated detector.

In the second part of the experiment, the percentage of false positives was varied. Due to the fact that the positions of real false positives are strongly correlated with certain background objects and with true pedestrians, real false positives generated by the Poselets detector were used rather than randomly scattered detections. False positives were identified as Poselets detections overlapping by less than 50% with the ground-truth [6]. At each frame, a threshold based on detector confidence was used to control the percentage of the total number of false positives added to the ground-truth detections. The results in Fig. 7 (b), show that addition of false positives causes a near linear decrease in tracker performance. This shows the tracker is sensitive to distractors, which justifies the pre-processing steps, such as filtering by height and detector confidence, taken before passing the raw output from the pedestrian detector to the tracker. These findings are similar to those of [8], included in Fig. 7, which also found that false positives are more detrimental to tracker performance than missed detections. Realistic tracker performance results given known pedestrain detector accuracy from [3], are also included in Fig 7 for reference.

E. Literature Comparison using Public Detection Sets

From the results of the previous experiments (See Section III-D) it is clear that tracker performance is highly correlated with detector performance. Therefore, to allow comparison of our tracker's performance with systems from the literature, we evaluated its performance using publicly available detection sets. For this experiment the PETS 2009 S2.L1 sequence was used due to the public availability of several widely used detection sets. For all detection sets, all tracker settings remained identical to those used when evaluating the tracker using the Poselets detector. The results from this experiment are shown in Table II.

Detection Set	Method	MOTA	MOTP	prec.	recall
B.Yang [59]	Whole System	91.73	69.13	96.54	95.14
	1st Stage Only	90.02	70.62	97.92	91.97
A.Andriyenko [4]	Whole System	88.91	78.12	95.98	92.80
	1st Stage Only	85.57	78.40	96.55	88.74
Poselet [9]	Whole System	87.9	72.5	95.3	92.5
	1st Stage Only	85.6	72.6	95.4	90.0
Ground Truth	Whole System	98.37	70.53	99.30	99.07
	1st Stage Only	98.35	70.53	99.28	99.07

TABLE II
TRACKING RESULTS PRODUCED BY SYSTEM, ON THE PETS S2.L1 SEQUENCE, USING PUBLICLY AVAILABLE DETECTION SETS.

These results further confirm that the detection set used is a major factor contributing to overall tracker performance, while also showing that our tracker is not overly dependent on, or tuned to a specific pedestrian detector. The value of the second stage tracklet linking process can be inferred by comparing the results for *1st Stage only* and *Whole System*, which correspond to the tracker used with and without tracklet linking respectively. Improved MOTA is observed for all detection sets when the tracklet linking mechanism is used. Tracking results produced using the ground-truth as a detection set have been included to illustrate the tracker's potential given a perfect detector. In this case, accuracy is still slightly degraded, due to the trade-offs necessary for dealing with the false positives and missed detections of a real detector.

By examining the MOTA results from the 3rd row of Table II we can see that given the same detection set, i.e. given equal conditions, our whole system was capable of outperforming the system of [4], which has a published MOTA of 81.4 and MOTP of 76.1, while obtaining marginally lower or comparable results to [5], which has a published MOTA of 89.3 and MOTP of 56.4. In addition, whereas the systems in [4] and [5] are based on global optimisation, our tracking framework is capable of working in real-time using the first-stage only, and of producing more accurate results at low-latency, by using the second stage to perform tracklet linking over a short sliding window. This implies a significant advantage to our proposed methodology for real time processing. Finally, although the MOTA score of [5] improves over [4], the MOTP score decreases. Therefore, it can be seen that, given the same detections, our system produces a comparable MOTA score, while having a higher MOTP score than [5], meaning that it more precisely follows the pedestrian paths. Note that the results of [59] were not published using the CLEAR MOT metrics, therefore we do not directly compare our results with this work. The results using these detections are included simply to illustrate the large variation in performance that can be caused by the detection set used.

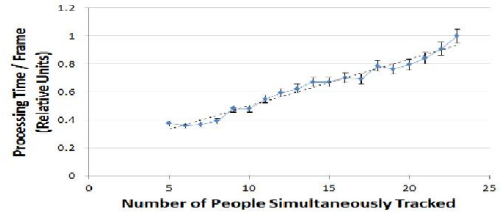


Fig. 8. The relative time to process each frame as a function of the number of people being simultaneously tracked. Also shown is a line fitted with linear regression to the data points.

F. Tracker efficiency

To assess the tracker's computation efficiency, a run was performed on the Oxford sequence. At each frame, the number of people tracked, along with processing time, was recorded. A linear regression line was fitted to this data, and for each number of persons, the mean and variance of processing time was calculated.

From results of this experiment, shown in Fig. 8, we can see there is a near linear relationship between the number of people tracked and the time to process each frame. The time to process each frame also exhibits low variance. Note that, time is shown in relative units as the goal of this experiment is to investigate relative efficiency, rather than absolute performance. The results indicate that the tracking methodology used is efficiently scalable to moderate numbers of people.

G. Comparison with the Literature

Shown in Table III are comparisons of the results from our tracking system with state-of-the-art trackers on the Oxford, and PETS sequences. The results from our system are broken down into those produced using the first stage only, and those produced using the whole tracking system, including optical flow tracklet linking, in order to show the relative contribution of each stage.

By examining the results in Table III, we can observe that the MOTA of our approach is better than, or comparable with most of the methods in the literature, even against techniques based on global optimisation [63], [28], [5], or including more complex reasoning, such as social behaviour [33], [46], [57]. For the Oxford dataset, we outperform an other registered method. As the performance of the first stage alone on the Oxford dataset is already very high, we observe a relatively small improvement with the addition of the second-stage tracklet linking mechanism. We have observed that, for this sequence, the first stage does not produce very fragmented tracklets, and many tracklets end at the edge of the frame, therefore there are relatively few opportunities where the second stage could potentially improve tracking performance. The high performance of the first stage is likely to be due to this dataset's high-resolution, meaning the Poselets pedestrian detector can perform very well, and the relatively linear motion of most of the pedestrians, meaning the predictions of the linear motion model are well matched to this scenario. This reinforces our decision to include a prediction/estimation mechanism within the LAP tracking framework.

On the contrary, with the PETS sequences, due to erratic and unpredictable motion trajectories of the subjects and total occlusions due to static objects, there are more occasions where tracklet linking can occur. The results show that tracklet linking can induce performance gains compared to the first stage used alone, particularly for the S2.L1 sequence. For the S2.L1 sequence, our results are comparable with approaches such as [5], [4], [3], but the results from the global optimisation approach of [23] exceed those of our system. Nonetheless, unlike [23] our approach has the advantage that it can produce results in near real-time, using a short sliding-window of frames. Our method also exhibits less sensitivity to parameter settings than [23]. On the S2.L1 sequence, our approach is out-performed by [62], while in S2.L3 and Oxford town centre sequence our method has higher MOTA. Likewise, we cannot properly evaluate performance against [60], as results are only provided for the S2.L1 sequence. The PETS S2.L2 and S2.L3 sequences are extremely challenging, due to high crowd density. On these sequences, our method is comparable with [3], [22], and again only significantly exceeded by the global optimisation method of [23]. However, on the S2.L2 and S2.L3 sequences our system has a higher MOTP value than the compared methods, meaning it is better able to accurately track the position of persons, even in these extremely challenging crowded conditions.

Example frames showing the output of the tracker on, PETS 2009 S2.L1 and the Oxford Town Centre dataset are shown in Fig. 9. As can be seen in both sequences, the majority of pedestrians are accurately tracked, with only a small number of errors.

IV. Conclusions

In this paper we have introduced a novel online dual-stage multi-target tracking framework, that is capable of handling partial occlusions and of linking broken tracks.

The system includes a novel occlusion-robust method for calculating the appearance similarity between tracks and detections, that does not require explicit identification of the occluded regions. This novel occlusion-robust appearance similarity method is used in the first stage of the tracker to produce short reliable tracklets, and in the second stage to link tracklets based on appearance. We shown that in realistic tracking scenarios this occlusion robust appearance similarity has the potential to more reliably associate tracks with correct detections, in the presence of distractors, compared to conventional appearance similarity methods.

In the second stage we have demonstrated a novel tracklet linking mechanism, that uses complementary features to constrain the linking problem and thus removes the need for ad-hoc rules governing the linking process. This was achieved by modelling the motion of each tracklet using optical flow features, so that only those links found to be compatible with

the optical flow evidence were permitted. We have shown that, given both simulated and real detection-sets from a variety of public sources, this method was able to consistently improve tracking performance under realistic conditions.

REFERENCES

[1] Ieee intl workshop on performance evaluation of tracking and surveillance (PETS). http://www.cvg.rdg.ac.uk/PETS2009/, 2009.

[2] E. Andrade, S. Blunsden, and R. Fisher. Characterisation of optical flow anomalies in pedestrian traffic. In *The IEEE Intl Symposium on Imaging for Crime Detection and Prevention*, pages 73–78, 2005.

[3] A. Andriyenko, S. Roth, and K. Schindler. An analytical formulation of global occlusion reasoning for multi-target tracking. In *IEEE International Conference on Computer Vision Workshops*, pages 1839–1846, Nov 2011.

[4] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *IEEE Conf on CVPR*, pages 1265–1272, 2011.

[5] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *IEEE Conf on CVPR*, 2012.

[6] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *IEEE Conf on CVPR*, 2011.

[7] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *PETS-Winter*, 2009.

[8] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Trans on PAMI*, 33(9):1806–1819, 2011.

[9] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Intl Conf on Computer Vision*, 2009.

[10] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Intl Conf on Computer Vision*, 2009.

[11] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Trans on PAMI*, 33(9):1820–1833, 2011.

[12] A. A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *IEEE Conf on CVPR*, 2013.

[13] Y. Cai, N. de Freitas, and J. Little. Robust visual tracking for multiple targets. In *European Conf on Computer Vision*, 2006.

[14] I. Cox and S. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Trans on PAMI*, 18(2):138–150, 1996.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf on CVPR*, 2005.

[16] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *IEEE Conf on CVPR*, 2010.

[17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans on PAMI*, 32(9):1627–1645, 2009.

[18] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans on PAMI*, 30(2):267–282, 2008.

[19] A. P. French, A. Naeem, I. L. Dryden, and T. P. Pridmore. Using social effects to guide tracking in complex scenes. In *IEEE Conf on Advanced Video and Signal Based Surveillance*, pages 212–217, 2007.

[20] J. Garcia, A. Gardel, I. Bravo, J. Lazaro, and M. Martinez. Tracking people motion based on extended condensation algorithm. *IEEE Trans on Systems, Man, and Cybernetics: Systems*, 43(3):606–618, 2013.

[21] C. Gong, K. Fu, A. Loza, Q. Wu, J. Liu, and J. Yang. Pagerank tracker: From ranking to tracking. *IEEE Trans on Cybernetics*, (99):1–1, 2013.

[22] M. Hofmann, M. Haag, and G. Rigoll. Unified hierarchical multi-object tracking using global data association. In *IEEE Intl Workshop on PETS*, pages 22–28, 2013.

[23] M. Hofmann, D. Wolf, and G. Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3650–3657, June 2013.

[24] C. Huang, Y. Li, and R. Nevatia. Multiple target tracking by learning-based hierarchical association of detection responses. *IEEE Trans on PAMI*, 35(4):898–910, 2013.

[25] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Computer Vision–ECCV 2008*, pages 788–801. 2008.

[26] N. Ihaddadene and C. Djeraba. Real-time crowd motion analysis. In *Intl Conf on Pattern Recognition*, pages 1–4, 2008.

[27] S. Iwase and H. Saito. Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *IEEE Intl Conf on Pattern Recognition*, pages 751–754, 2004.

[28] H. Izadinia, I. Saleemi, W. Li, and M. Shah. (mp)2t: multiple people multiple parts tracker. In *European Conf on Computer Vision*, 2012.

[29] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser. Robust single-particle tracking in live-cell time- lapse sequences. *Nature methods*, 5(8):695–702, 2008.

[30] B. Keni and S. Rainer. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008.

[31] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *European Conf on Computer Vision*, 2004.

[32] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[33] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Intl Conf on Computer Vision*, 2011.

[34] J. Liu, P. Carr, R. T. Collins, and Y. Liu. Tracking sports players with context-conditioned motion models. In *IEEE Conf on CVPR*, pages 1830–1837, 2013.

[35] M. Luber, J. Stork, G. Tipaldi, and K. Arras. People tracking with human motion predictions from social forces. In *IEEE Intl Conf on Robotics and Automation*, pages 464–469, 2010.

[36] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *Intl Journal of Computer Vision*, 19(1):57–71, 2000.

[37] V. Maroulas and P. Stinis. Improved particle filters for multi-target tracking. *Journal of Computational Physics*, 231(2):602 – 611, 2012.

[38] N. McLaughlin, J. M. Del Rincon, and P. Miller. Online multiperson tracking with occlusion reasoning and unsupervised track motion model. In *IEEE Intl Conf on Advanced Video and Signal Based Surveillance*, pages 37–42, 2013.

[39] N. McLaughlin, J. Ming, and D. Crookes. Robust bimodal person identification using face and speech with limited training data and corruption of both modalities. In *Interspeech*, 2011.

[40] J. Ming, T. Hazen, J. Glass, and D. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Trans on Audio Speech and Language Processing*, 15(5):1711–1723, 2007.

[41] A. Mittal and L. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Computer Vision ECCV 2002*, volume 2350 of *Lecture Notes in Computer Science*, pages 18–33. Springer Berlin Heidelberg, 2002.

[42] V. Nagarajan, M. Chidambara, and R. Sharma. Combinatorial problems in multitarget tracking - a comprehensive solution. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, volume 134, pages 113–118, 1987.

[43] K. Okuma, A. Taleghani, N. D. Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conf on Computer Vision*, 2004.

[44] C. Otto, W. Gerber, F. Leon, and J. Wirnitzer. A joint integrated probabilistic data association filter for pedestrian tracking across blind regions using monocular camera and radar. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 636–641, 2012.

[45] H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors. In *International Joint Conference on Artificial Intelligence*, volume 2, pages 1160–1167, 1999.

[46] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE Computer Vision*, 2009.

[47] Z. Qin. Improving multi-target tracking via social grouping. In *IEEE Conf on CVPR*, pages 1972–1978, 2012.

[48] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Trans on Automatic Control*, 24:843–854, 1979.

[49] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *IEEE Intl Conf on Computer Vision*, pages 1389–1396, 2009.

[50] G. Rogez, J. Rihan, J. Guerrero, and C. Orrite. Monocular 3-d gait tracking in surveillance scenes. *IEEE Transactions on Cybernetics*, (99):1–1, 2013.

[51] K. Shafique and M. Shah. A noniterative greedy algorithm for multi-frame point correspondence. *IEEE Trans on PAMI*, 27(1):51–65, 2005.

[52] G. Shu, A. Dehghan, and O. Oreifej. Part-based multiple-person tracking with partial occlusion handling. In *IEEE Conf on CVPR*, 2012.

[53] V. Singh, B. Wu, and R. Nevatia. Pedestrian tracking by associating tracklets using detection residuals. In *IEEE Workshop on Motion and video Computing*, pages 1–8, 2008.

[54] P. M. Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489–490, 2008.

[55] L. Wang, S. Jia, X. Li, and S. Wang. Human gait recognition based on gait flow image considering walking direction. In *Intl Conf on Mechatronics and Automation*, pages 1990–1995, 2012.

Sequence	Method	MOTA	MOTP	Precision	Recall
PETS S2.L1	Our Method (Whole System)	87.9	72.5	95.3	92.5
	Our Method (1^{st} Stage Only)	85.6	72.6	95.4	90.0
	J. Berclaz *et al. [8]*	68.4	63.3	79.1	90.3
	A. Andriyenko *et al. [4]*	81.4	76.1	-	-
	A. Andriyenko *et al. [5]*	89.3	56.4	-	-
	A. Andriyenko *et al. [3]*	88.3	75.7	-	-
	M. Breitenstein *et al. [11]*	79.7	56.3	-	-
	M. Hofmann *et al. [23]*	98.0	82.8	-	-
	M. Hofmann *et al. [22]*	97.8	75.3	-	-
	B. Yang *et al. [59]*	-	-	99.0	91.8
	J. Zhang *et al. [62]*	93.3	68.2	-	-
	Y. Yi *et al. [60]*	94.8	85.5	-	-
PETS S2.L2	Our Method (Whole System)	57.5	72.8	88.4	66.3
	Our Method (1^{st} Stage Only)	56.8	73.0	90.4	63.6
	A. Andriyenko *et al. [3]*	60.2	60.5	-	-
	M. Hofmann *et al. [23]*	75.8	72.1	-	-
	M. Hofmann *et al. [22]*	57.1	56.4	-	-
	J. Zhang *et al. [62]*	66.7	58.2	-	-
PETS S2.L3	Our Method (Whole System)	53.6	74.1	87.5	62.5
	Our Method (1^{st} Stage Only)	53.3	74.1	87.7	62.0
	A. Andriyenko *et al. [3]*	43.8	66.3	-	-
	M. Hofmann *et al. [23]*	62.8	70.5	-	-
	M. Hofmann *et al. [22]*	41.5	65.0	-	-
	J. Zhang *et al. [62]*	40.4	56.4	-	-
Oxford Town Centre	Our Method (Whole System)	80.8	70.9	95.3	85.2
	Our Method (1^{st} Stage Only)	80.7	70.9	95.4	85.1
	H. Izadinia *et al. [28]*	75.7	71.6	93.6	81.8
	B. Benfold *et al. [6]*	61.3	80.3	82.0	79.0
	G. Shu *et al. [52]*	72.9	71.3	-	-
	K. Yamaguchi *et al. [57]*	61.3	70.9	71.1	64.0
	S. Pellegrini *et al. [46]*	63.4	70.7	70.8	64.1
	L. Zhang *et al. [63]*	65.7	71.5	71.5	66.1
	L. Leal-Taixe *et al. [33]*	67.3	71.5	71.6	67.6
	J. Zhang *et al. [62]*	73.6	68.8	-	-

TABLE III
COMPARISON OF THE RESULTS PRODUCED BY OUR SYSTEM WITH THE LITERATURE FOR THE OXFORD AND PETS DATASETS.

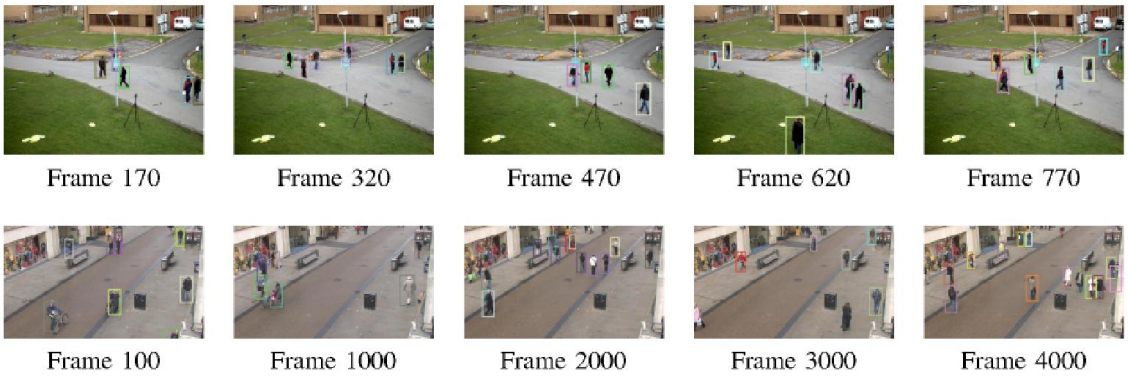


Fig. 9. Output of the complete tracking system, in the top row on tested on the PETS 2009 S2.L1 sequence (every 150 frames), and in the bottom row tested on the Oxford Town Centre dataset (every 1000 frames). Full videos are available in the supplemental material of this paper.

[56] J. Wolf, A. Viterbi, and G. Dixon. Finding the best set of k paths through a trellis with application to multitarget tracking. *IEEE Trans on Aerospace and Electronic Systems*, 25(2):287–296, 1989.
[57] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *IEEE Conf on CVPR*, 2012.
[58] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a crf model. In *IEEE Conf on CVPR*, 2011.
[59] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *IEEE Conf on CVPR*, pages 1918–1925, 2012.
[60] Y. Yi and H. Xu. Hierarchical data association framework with occlusion handling for multiple targets tracking. *IEEE Signal Processing Letters*, 21(3):288–291, March 2014.
[61] J. Zhang, L. Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance*, pages 379–385, Sept 2012.
[62] J. Zhang, L. Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 379–385, Sept 2012.
[63] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *IEEE Conf on CVPR*, 2008.